Документирование языков с использованием среды SIL FieldWorks Language Explorer (на материале осетинского языка)[1]

Беляев О. И.
(МГУ им. М. В. Ломоносова)
Выдрин А. П.
(ИЛИ РАН)

На примере разработанного в рамках среды SIL FieldWorks морфологического описания осетинского языка, использованного для создания онлайн-корпуса отглоссированных текстов, рассматриваются новые возможности, позволяющие ускорить и сделать более доступным процесс языкового документирования и сбора языковых данных.

Based on the morphological description of Ossetic developed using SIL FieldWorks which has been used for creating an online corpus of interlinear texts, the aim of this paper is to discuss the new possibilities which allow to make the process of language documentation and data gathering faster and more productive.

0. Введение

В повседневной практике полевой лингвистики исследователи используют различные средства, призванные облегчить работу с языковыми данными. В области сбора словаря и анализа текстов фактическим стандартом уже стало приложение Toolbox (известное в более ранних версиях как Shoebox), позволяющее в достаточно гибком, стандартизированном формате представлять как лексическую, так и текстовую информацию. Важной возможностью среды Toolbox является возможность автоматического глоссирования (interlinearization) вводимых текстов на основании словаря и заданных пользователем правил соединения морфем в слова. О применении Toolbox для создания корпусов отглоссированных текстов см., к примеру, Кибрик et al. 2007. В качестве других недавних примеров использования Toolbox для автоматического глоссирования текстов и полевых исследований можно привести работы Robinson, Aumann and Bird 2007 и McGill 2009.

Несмотря на то, что среда Toolbox, несомненно, сильно облегчает работу лингвиста, она имеет ряд существенных недостатков. Прежде всего, интерфейс программы не всегда интуитивен и требует от пользователя определённых технических навыков. Гибкость программы достигается за счёт максимальной абстракции от типа задаваемых данных. В связи с этим в SIL был разработан новый программный пакет под названием FieldWorks, одной из составных частей которого является FieldWorks Language Explorer (FLEx). FLEx аналогичен Toolbox и способен заменить его в том смысле, что также предоставляет возможности работы со словарём, задания морфологических правил и автоматического глоссирования текстов. При этом по сравнению с Toolbox FLEx обладает тем преимуществом, что в большей степени ориентирован на полевых лингвистов, не имеющих опыта работы с формальными компьютерными моделями языка, и использует привычные для них термины; кроме того, возможности FLEx в плане составления словаря и автоматической обработки текстов значительно превосходят возможности,

[1] Исследование проводилось при частичной финансовой поддержке гранта РГНФ 09-04-00168а.

предоставляемые Toolbox. Если в Toolbox имеется только два основных компонента – словарь и тексты, а возможность задавать грамматику существует лишь в форме линейных правил, то FLEx содержит отдельный самостоятельный грамматический компонент, форма представления правил в котором близка к стандартам, принятым в современных лингвистических описаниях. Кроме того, важнейшим преимуществом FLEx перед Toolbox является то, что любые изменения, внесённые в словарь, автоматически отражаются в ранее отглоссированных текстах.

В ходе работы над проектом документирования осетинского языка мы успешно совершили переход от использования Toolbox к использованию FLEx. Благодаря использованию FLEx нам удалось в короткие сроки создать постоянно обновляемый мини-корпус отглоссированных осетинских текстов, доступный в Интернете по адресу http://pole.iphil.ru/ossetic . По нашему мнению, FLEx способен полностью заменить Toolbox как средство глоссирования текстов, а также использоваться для сбора словаря вместо часто используемых в этих целях самостоятельных утилит. Основная часть статьи будет посвящена нашему опыту использования FLEx при документировании осетинского языка, а также возникшим при этом трудностях и путям их решения.

1. Словарь

Словарь позволяет не только отобразить саму словарную единицу [2] с морфологическими пометами, алломорфами, вариантами и дериватами, переводом, примерами и заметками, но и снабдить её произношением и иллюстрацией. Пример словарной карточки приведён на рис. 1.

[2] В терминологии FLEx, «лексему» (lexeme) – однако во избежание омонимии мы будем называть «лексему» FLEx «словарной единицей», а термин «лексема» использовать только в его лингвистическом смысле, поскольку словарные единицы FLEx включают в себя как корневые морфемы, так и аффиксы и даже словоформы.

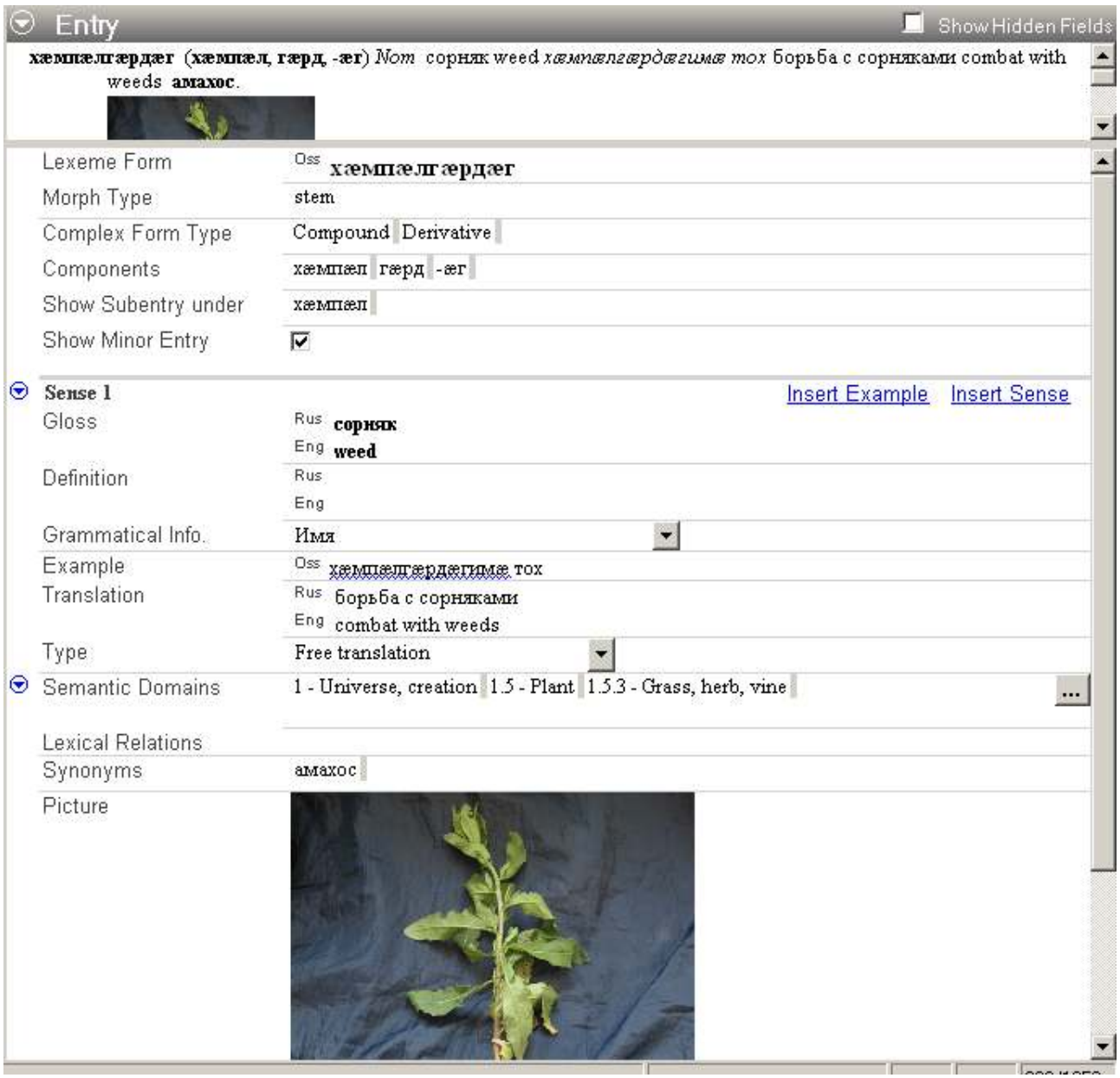

Рис. 1. Пример словарной карточки

Как можно видеть на Рис. 1, в верхней части словарной карточки отображается словарная статья, то есть конечный вид лексической единицы при публикации словаря. К сожалению, в данной версии FLEx возможности по форматированию словарной статьи достаточно ограничены, однако в перспективе эта функция должна позволить получать на выходе словарь, готовый к публикации. В таком виде словарь отображается при выборе режима Dictionary в левой панели программы.

Общее число полей существенно больше отображаемых на Рис. 1, что даёт возможность пользователю выбирать, какая степень детализации подачи информации ему необходима.

С точки зрения анализа текстов существенное преимущество FLEx состоит в том, что в словарной карточке отображается также грамматическая информация, позволяющая отдельно не создавать отдельный набор правил для автоматического снятия омонимии (как это было ранее в программе Toolbox). Например, ранее, при использовании программы Toolbox, для осетинского языка (как и для любых языков с несколькими основами для прошедшего и настоящего времени и с частично совпадающим набором

парадигм презенса и претерита) приходилось создавать формулы вроде v.pst (morph.v.pst) и v.prs (morph.v.prs), а каждой морфеме, таким образом, приписывать в словаре особые «части речи» (лишь в качестве технического решения), соответствующие тем частям речи, к которым эти морфемы присоединяются. В настоящей программе словарь коррелирует с грамматикой, в которой прописываются принципы сочетаемости морфем друг с другом (об этом см. ниже). Такой метод представления, возможно, имеет свои недостатки, однако с лингвистической точки зрения значительно ближе к реальной грамматике языка, чем метод, представленный в Toolbox.

Кроме того, значительное преимущество FLEx состоит в том, что любое изменение, внесённое в любое из полей словарной статьи (например, в поле глоссы), автоматически отразится на всех ранее отглоссированных текстах.

При использовании словарного компонента программы мы столкнулись со следующими проблемами:

1. При создании алломорфов невозможно создавать для них отдельные глоссы. С технической точки зрения такое ограничение оправдано: в то время как алломорфы относятся ко всей словарной единице целиком, глоссы относятся к отдельным её значениям. Однако оно создаёт проблемы при наличии у лексемы нескольких основ. В нашем случае это основы настоящего и прошедшего времени глагола (см. Абаев 1970: 594), которые распределены морфологически, то есть присоединяют разные окончания и не несут собственной грамматической информации. В то же время для удобства восприятия тип основы принято всегда глоссировать (например, *кæн-ын* ‘делать:PRS-PRS.1SG’). Единственным возможным решением в рамках FLEx могло бы быть разделение двух основ по разным лексемам. Однако такое решение нецелесообразно в силу своей громоздкости (создаётся, к примеру, дополнительная проблема связывания двух лексем между собой) и отсутствия какой бы то ни было лингвистической мотивации[3]. Адекватного решения данной проблемы нами пока не найдено.
2. Не имеет адекватного решения проблема супплетивных словоформ и фузии. На данный момент специальных средств для такого рода случаев FLEx не предусматривает, в связи с чем супплетивные словоформы приходится вводить в словарь как отдельные словарные единицы. Это решение не вполне идеально, поскольку хотелось бы, чтобы словарным единицам соответствовали корневые и аффиксальные морфемы, а не целые словоформы. Супплетивные же формы было бы удобно описывать внутри той словарной единицы, к которой они относятся (то есть, например, при формализации парадигмы местоимения *æз* ‘я’ приходится создавать словарные единицы практически для каждой его падежной формы, тогда как желательно было бы поместить все эти формы в одно гнездо с номинативом *æз*). Проблема фузии является ещё более сложной и будет рассмотрена в разделе «Грамматика».
3. Понятие «варианта словарной единицы» является, тем самым, достаточно

[3] Заметим, что для некоторых других языков, например, даргинских или новогреческого, такое решение, напротив, является правильным, так как в этих языках основа имеет аспектуальное значение, и две основы могут присоединять одни и те же наборы окончаний. В осетинском же распределение основ имеет чисто морфологическую природу, т. е., например, от основы «настоящего времени» образуются также будущее время и конъюнктив, а от основы «прошедшего времени» - контрафактив, и все эти формы имеют разные наборы окончаний.

размытым: «вариантами» оказываются как диалектные/социолектные формы лексем, так и супплетивные формы, а также (если избрать такой метод описания) – нерегулярно образующиеся формы, например, множественного числа или прошедшего времени. Желательно было бы отделить понятие собственно «варианта» от понятия нерегулярной формы.

В целом, впрочем, следует признать, что словарь в FLEx существенно улучшен по сравнению с Toolbox и удобен в использовании, во всяком случае для наших задач. Возможно, для лексикографических целей средств, предоставляемых FLEx, недостаточно, и необходимо использование специально разработанных для этого утилит, однако на этапе создания словаря для глоссирования текстов полевому лингвисту несомненно будет достаточно технических средств FLEx.

2. Грамматика

Грамматический компонент FLEx существенно отличается от Toolbox, где грамматика описывается посредством линейных правил, и поэтому сравнивать FLEx с более ранними приложениями здесь не представляется осмысленным. Грамматическая информация, которой пользователь снабжает FLEx, обрабатывается встроенным парсером при автоматическом анализе текстов.

Грамматическая часть FLEx тесно связана с лексической. Грамматика как таковая задаётся как иерархически организованный набор частей речи (лексических категорий). Для каждой из этих частей речи существуют модели словоформы, по которым изменяются их лексемы (Affix templates). Каждая из моделей словоформы состоит из линейно упорядоченных ячеек (слотов, Slots), куда могут помещаться различные аффиксы. Слоты могут быть обязательными или необязательными (optional). Информация о том, к какой категории присоединяется аффикс и какой слот он занимает, содержится в словаре. Пример основной части грамматического интерфейса FLEx, режима редактирования частей речи (Category Edit), можно видеть на Рис. 2.

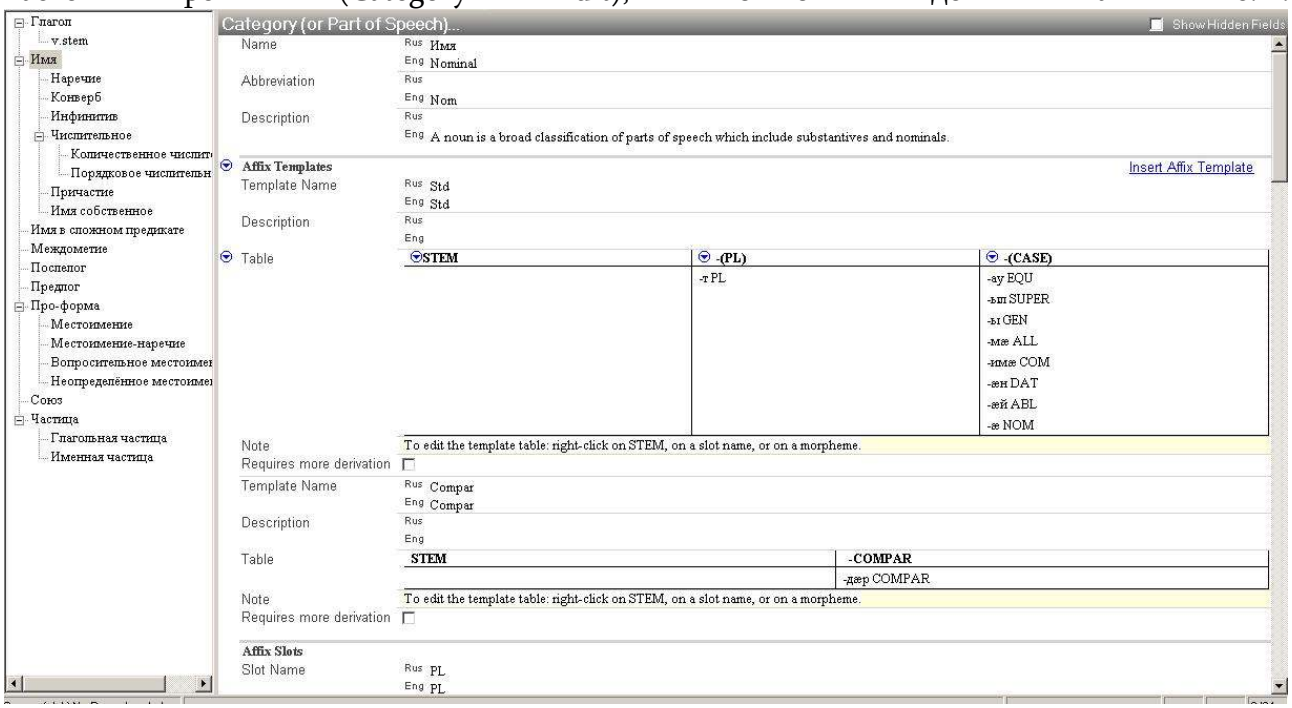

Рис. 2. Режим редактирования частей речи

Помимо возможности задавать модели словоформ в виде набора линейно упорядоченных слотов, грамматическая система FLEx предоставляет в распоряжение пользователя систему фонологических правил (Phonological rules), фонетических окружений (environments), используемых для того, чтобы ограничить сочетаемость алломорфов, а также систему признаков (features). Фонологические правила были добавлены в последней версии FLEx и в теории позволяют задавать для цепочек фонем трансформационные правила в духе классической генеративной фонологии (см. Chomsky, Halle 1968). Однако эта возможность пока что носит экспериментальный характер и к тому же требует использования отдельного парсера, не поддерживающего многие другие возможности FLEx, такие как систему признаков. В связи с этим нашим проектом фонологические правила пока что не используются.

Фонетические окружения позволяют ограничить встречаемость алломорфов. Они задаются в виде строки символов, в которой можно использовать, помимо символа «текущий элемент» (_) и обозначений фонем, символы начала слова, конца слова, а также т. н. «естественные классы» (Natural classes) фонем, например «гласные», «согласные», «абруптивы», которые также задаются пользователем. Например, если для алломорфа *–йы* морфемы генитива *–ы* добавить окружение, задаваемое строкой «/[V]_», то это ограничит использование этого алломорфа до позиции после гласных, что и наблюдается в осетинском языке (см., например, Ахвледиани 1963: 90). Впрочем, для осетинского языка здесь возникает та трудность, что в принятой орфографии, используемой и нами в наших текстах, знак *у* обозначает как гласный /u/, так и полугласный /w/. Таким образом, он должен быть отнесён одновременно и к гласным и к согласным. Решением проблемы, хотя и несколько громоздким, является создание дополнительных окружений специально для этой фонемы, вроде «/[C]у_», т. е. требованием нахождения «у» в позиции перед согласной.

Что касается механизма признаков, то он весьма удобен при описании морфологии языка и широко применяется нами при описании осетинского языка. Механизм признаков позволяет одним морфемам требовать присутствия других морфем, которым приписан тот или иной «признак» (который может и не иметь лингвистического содержания, а носить чисто технический характер). Механизм признаков включает в себя использование «имён основ» (Stem Names) – каждая из основ задаётся как набор признаков, который должен присутствовать в присоединяемых к ней морфемах.

В нашем проекте «имена основ» используются для основ настоящего и прошедшего времени глагола; у имён – для основы множественного числа (где часто происходит чередование гласных, см. Багаев 1965: 130), а также для отдельной «основы» для форм, где происходит палатализация конечного заднеязычного согласного перед падежным аффиксом *-ы* (генитив и инэссив, напр. *лæг* ‘человек’ – *læджы* ‘человека’). В последнем случае мы могли бы вместо использования отдельной основы задать фонетическое окружение вроде «¥_ы», или даже фонетическое правило для палатализации, однако этому мешает два обстоятельства: во-первых, палатализации не происходит перед окончанием суперэссива, также начинающимся на /ы/ (ср. *лæгыл* ‘на человеке’); во-вторых, если бы это чередование было чисто автоматическим, тогда оно происходило бы и перед *–и*, однако при присоединении окончания комитатива, начинающегося на этот гласный, палатализации не происходит (ср. *лæгимæ* ‘с человеком’). Таким образом, несмотря на частотность этого чередования в осетинском языке, техническая невозможность описать его как фонетически обусловленное служит доводом в пользу того, что оно является

морфонологическим[4].

При работе с грамматической частью FLEx мы столкнулись со следующими проблемами:

1. Проблема фузии морфем, о которой уже было упомянуто выше. Морфологическая модель, представляющая словоформу как последовательность линейно упорядоченных «слотов» для аффиксом, не приспособлена для случаев фузии без использования специальных механизмов. В последних версиях FLEx такие механизмы, по всей видимости, отсутствуют. Для осетинского языка характерно два проблемных случая фузии:
    a. Соединение преверба с глагольной основой, напр. *фæ + уынын* ‘видеть’ – *фенын* ‘увидеть’. В этом случае возможно два решения: либо считать превербом *фе-* (такой алломорф у *фæ-* существует в любом случае), а основой настоящего времени – согласный *-н-* (*-ын* – суффикс инфинитива). Тогда основой прошедшего времени будет *-д-,* ср. *фе-д-т-он* «PREF-видеть:PST-TR-1SG» ‘я увидел’. Это решение очевидным образом искусственно и к тому же создаёт множество потенциальных проблем для парсера. Единственный его плюс – это сохранение «слотового» принципа организации словоформы, но в данном случае это не так существенно. Более правильным, на наш взгляд, решением является создание отдельной словарной единицы *фен-/фед-* с глоссой, в которой фузия отмечена знаком «+»: ‘PREF+видеть’.

       Слот преверба здесь оказывается незаполненным, и теоретически парсер может счесть правильной форму вроде **бафедтон* с двумя превербами (что невозможно в осетинском). Однако в реальных текстах такие формы не встречаются, и потому принятое нами решение можно считать удовлетворительным, хотя и не идеальным с лингвистической точки зрения.
    b. Случаи выпадения показателя переходности *-т*. Особенностью осетинского спряжения является различие парадигм прошедшего времени и контрфактива у переходных и непереходных глаголов (Багаев 1965: 276-280). Помимо различных личных окончаний в прошедшем времени, в переходном спряжении после основы добавляется показатель *-т*. В связи с тем, что основа прошедшего времени всегда оканчивается на *-д/-т*, это вызывает геминацию. Однако геминаты в осетинском языке могут выступать только после гласных или после сонорных согласных; во всех остальных случаях геминация снимается (Абаев 1949: 621). У большинства глаголов с основой прошедшего времени с исходом на шумный + /т/ аффикс переходности полностью выпадает и не оставляет никаких следов (ср. **фыст-т-он > фыстон*). Однако у глаголов с основой прош. вр. с исходом на шумный + /д/ показатель *-т-* вызывает оглушение (ср. **сыгъд-т-он > сыхтон* ‘я жёг’).

[4] Этому факту можно было бы предложить историческое объяснение: дело в том, что генитив и инессив в осетинском языке являются древними падежами, унаследованными от праиндоевропейского языка, тогда как и суперэссив, и комитатив являются новыми падежами, образованными от послелогов (ср. Камболов 2006). Однако это объяснение не действует на синхронном уровне, т. к. падежные маркеры никак не отличаются друг от друга в плане отделимости/переместимости.

Для большинства глаголов это в принципе не является проблемой, т. к. переходность в общем случае является словоклассифицирующей категорией и не различает грамматических форм. К тому же в прошедшем времени индикатива, помимо аффикса *-т-,* переходные глаголы присоединяют совершенно различные наборы окончаний. Проблемы возникают у небольшой группы «лабильных» глаголов с исходом основы прош. вр. на *–д* в формах контрфактива, где встречаются минимальные пары (**сыгъд-т-аид* > *сыхтаид* 'если бы он жёг' vs *сыгъд-аид* 'если бы он горел'). Правильным решением здесь было бы в основе *сыхт-* видеть фузию *сыгъд* и маркера *-т-*, однако система FLEx не даёт такой возможности. Поэтому приходится считать, что у глагола 'жечь' основой прош. вр. является *сыхт-*, а также что он требует признака [+transitive] – это позволяет избежать того, чтобы формы типа **сыгъд-тæн* > *сыхтæн* 'я горел' с окончанием непереходного типа (где /т/ входит в окончание) не считались формами глагола 'жечь'.

2. Проблема аффиксов с широкой сочетаемостью. В осетинском языке к ним относятся перфективирующие превербы, которые могут присоединяться как к глагольным словоформам, так и к именам, если они находятся в составе сложных предикатов. При этом, подобно русским глагольным приставкам, с лингвистической точки зрения превербы относятся скорее к области деривации, чем к области флексии, поскольку, помимо аспектуального значения, несут с собой также и нетривиальные семантические изменения.

   Однако сделать превербы деривирующими мешает то обстоятельство, что преверб всегда бывает только один (за исключением вторичного имперфектива -цæй-), а деривация позволила бы присоединять к глагольной словоформе бесконечное число превербов. Тогда, в частности, преверб *æрба-* разбирался бы как *æр-* + *ба-.* Такое функционирование грамматики нежелательно.

   Таким образом, наиболее оптимальным решением является создание для каждого преверба словарной единицы с двумя «значениями» (sense) – одно для имён, одно для глаголов. Уместить оба варианта употребления в одно значение позволило бы описание превербов как «проклитик» - FieldWorks рассматривает клитики как единицы, которые могут писаться как слитно, так и раздельно с их носителями. Однако как с фонологической точки зрения, так и с точки зрения отделимости превербы клитиками не являются, и потому нами было принято решение в пользу их многозначности.

3. Тексты

На Рис. 3 можно видеть общую форму разобранного текста в программе FLEx. Интерфейс FLEx позволяет совершать анализ на любом количестве языков. Предусмотрены следующие поля разбора словоформы: Morphemes (разделение фонетической формы слова на морфемы), Lexical Entries (представление морфологической структуры слова в виде исходных лексем, а не используемых алломорфов), Lexical Gloss (поморфемные глоссы), Lex. Gram. Info (информация о категории каждой из морфем; кроме того, для аффиксов – о слоте, а для корней – о типе склонения), Word Gloss (буквальный перевод слова), Word Cat (часть речи, к которой принадлежит слово целиком). Кроме того, для каждого предложения имеется буквальный и свободный перевод.

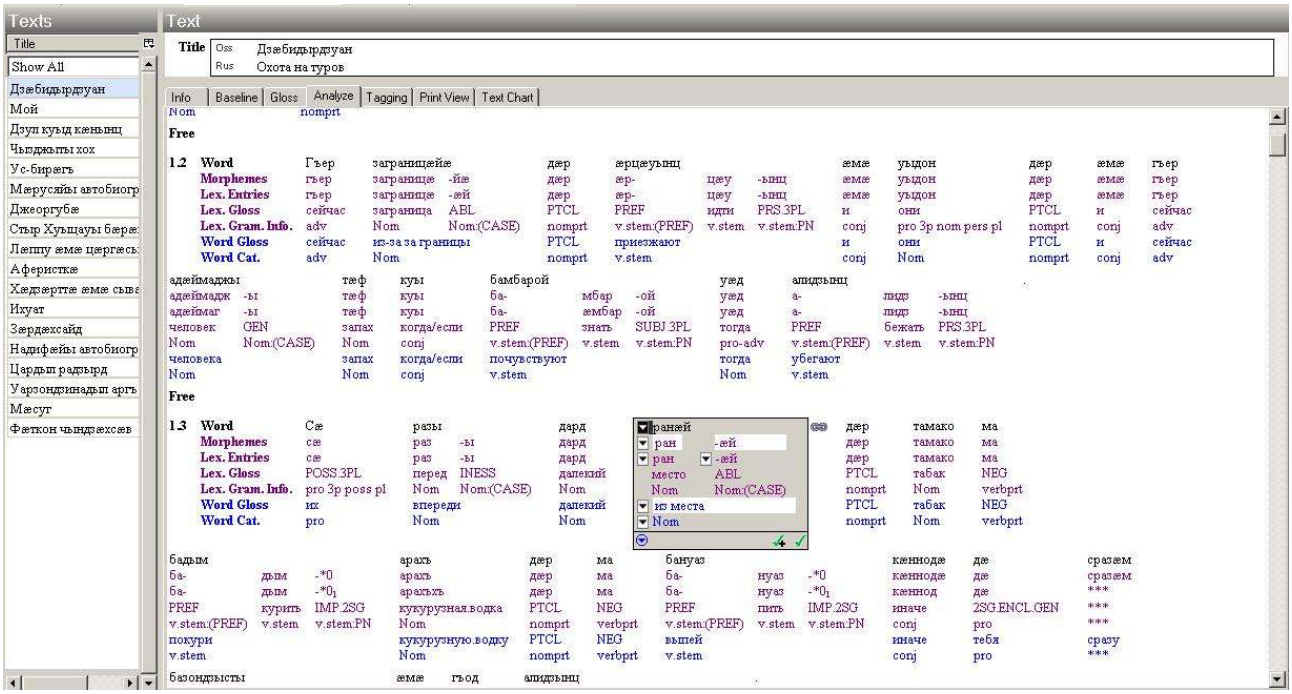

Рис 3. Общий вид отглоссированного текста

FLEx предусматривает возможность автоматического разбора словоформ в введённых текстах. Парсер руководствуется правилами, введёнными в разделе «грамматика», и в целом здесь можно повторить все те недостатки, которые были перечислены в разделе 2. Недостатком собственно парсера, а точнее интерфейса разбора текстов, является отсутствие возможности одновременного отображения нескольких разборов словоформ. В рамках программы возможно только два решения этой проблемы: 1. внесение в одну словарную карточку всех омонимов и последующие снятие омонимии в поле Word Gloss (см. пример на Рис. 4 ниже); 2. внесение омонимичных словарных карточек с последующим ручным выбором нужного омонима.

Word	мæ
Morphemes	POSS.1SG/1SG.ENCL.GEN/1SG.ENCL.ABL
Lex. Entries	мæ
Lex. Gloss	POSS.1SG/1SG.ENCL.GEN/1SG.ENCL.ABL
Lex. Gram. Info.	PREF/ENCL
Word Gloss	POSS.1SG
Word Cat.	PREF

Рис 4. Пример омонимичной словоформы

4. Корреляция между составляющими программы

Наиболее важной особенностью программы является автоматизированность корреляций между словарём и текстами, не требующая от пользователя внесения дополнительных настроек. При внесении изменений в любых полях словарной карточки, кроме Lexical entry (то есть фонетической формы слова), изменения автоматически отображаются во всех отглоссированных текстах. Это является существенным улучшением по сравнению с Toolbox, где приходилось в случае любых изменений в словаре заново запускать парсер.

5. Экспорт

В программе присутствует возможность экспорта в такие форматы, как OpenOffice.org (.odt), Microsoft Word 2003 XML, в собственный XML-формат, формат HTML и др. Благодаря возможности экспорта нами был в кратчайшие сроки разработан мини-корпус осетинских текстов, доступный по адресу http://pole.iphil.ru/ossetic/texts .

6. Заключение

В рамках нашего проекта нами был произведён переход от использования Toolbox при глоссировании текстов к использованию SIL FieldWorks Language Explorer (FLEx). Опыт можно признать удачным: в результате использования программы нами был в весьма короткие сроки разработан мини-корпус осетинских текстов, доступный для свободного доступа в интернете. Мы надеемся, что наш опыт и описание возникших при использовании программы проблем поможет будущим исследователям в их полевой практике.

Литература

1. Абаев В.И. Грамматический очерк осетинского языка // Орджоникидзе: 1959.
2. Ахвледиани Г.С. Грамматика осетинского языка // Орджоникидзе: 1963. Т. 1.
3. Камболов, Т.Т. Очерк истории осетинского языка // Владикавказ: Ир, 2006.
4. Кибрик А.Е., Архипов А.В., Даниэль М.А., Кодзасов С.В., Майерс Т., Нахимовский А.Д. Технологии обработки языковых данных в документировании малых языков //
5. Багаев Н.К. Современный осетинский язык // Орджоникидзе: Северо-Осетинское книжное издательство, 1965. Т. 1.
6. Chomsky N. and M. Halle. The sound pattern of Englis*h* // New York: 1968.
7. McGill Stuart. Documenting grammatical tone using Toolbox: an evaluation of Buseman's interlinearisation technique // Language Documentation and Description 6: 2009. С. 236-250.
8. Robinson S., G. Aumann and S. Bird. Managing Fieldwork Data with Toolbox and the Natural Language Toolkit // Language Documentation and Conservation 1: 2007. С. 44-57.